Formal Privacy and Synthetic Data for the American Community Survey

Michael H. Freiman, U.S. Census Bureau
Rolando A. Rodríguez, U.S. Census Bureau
Jerome P. Reiter, Duke University and U.S. Census Bureau
Amy Lauger, U.S. Census Bureau

FCSM Research and Policy Conference
March 8, 2018

The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

A roadmap for our research

1. Data protection goals for Census Bureau data
    - Protect data against all possible attacks
    - Quantify disclosure risk and data quality
2. Current methods do not fully achieve the goals
3. Differential privacy can achieve these goals
    - Assumed currently infeasible for American Community Survey (ACS)
4. Current synthesis research is intermediate step toward goals

The Census Bureau has multiple goals in protecting data

In the current data environment, we must protect data against known attacks and attacks not yet known to us

We need to quantify the risk and data quality that our disclosure metrics allow

Our methods should be transparent, so that users can account for the effect of disclosure on their inferences

Current methods to protect ACS respondents have shortcomings

ACS household data have historically been protected primarily by data swapping

Pairs of similar households have their geographic information switched

Some other methods are also used, particularly for the public use microdata

Some of the parameters of the disclosure protections are kept secret, preventing researchers from accurately adjusting their inferences

Methods cannot be mathematically demonstrated to be safe

Traditional disclosure avoidance methods are no longer adequate

Database reconstruction algorithms have become more effective

Computing power has increased

“Big data” means the Census Bureau can no longer assume it knows all of the data an attacker could use

Results of reidentification studies cannot generalize beyond particular datasets and attacks

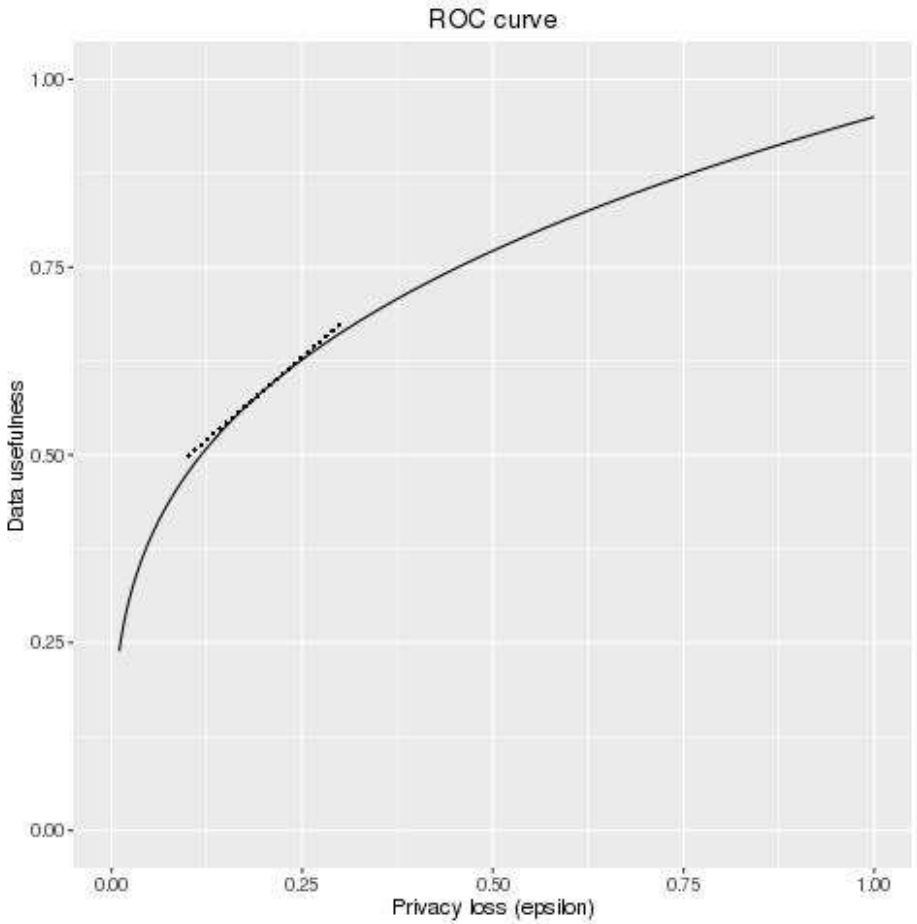
Formal privacy provides the guarantees traditional methods lack

A formal privacy framework, e.g., differential privacy, defines and quantifies the privacy loss from data releases

Algorithms used to protect the data must be proven to limit the privacy loss to no more than a certain “budget”

For the 2020 Census, the Census Bureau plans to create differentially private microdata, from which all data products will be derived

The ROC curve shows the tradeoff in setting the privacy budget



The choice of privacy budget is a tradeoff between data usefulness and privacy loss

More privacy requires more perturbation

More utility requires less privacy

The appropriate point on the curve is a subjective decision

Making the ACS formally private is particularly challenging

The ACS collects data from 2.3 million housing units and publishes 12 billion estimates per year

The ACS uses a complex sample weighting approach

Some challenges are shared with the census

- Statistics are desired for small geographic areas
- Some within-household relationships are important and should be reflected in the protected data

We aim to make the ACS formally private in the future, but that is not our current focus

Model-based synthetic data will improve on current methods but will not be formally private

We generate variables sequentially, not yet incorporating weights

Each variable in the synthesis is created using the previously synthesized variables for that record

We synthesize continuous variables using regression

We synthesize categorical variables with a classification tree

- Grow a tree on the previous variables
- Draw a value at random from the appropriate leaf

The order in which variables are synthesized considers the form order and the variable universes

The “universe” is the set of possible records for which a variable is defined

Some variables are defined for everyone

- Others are defined only for subpopulations

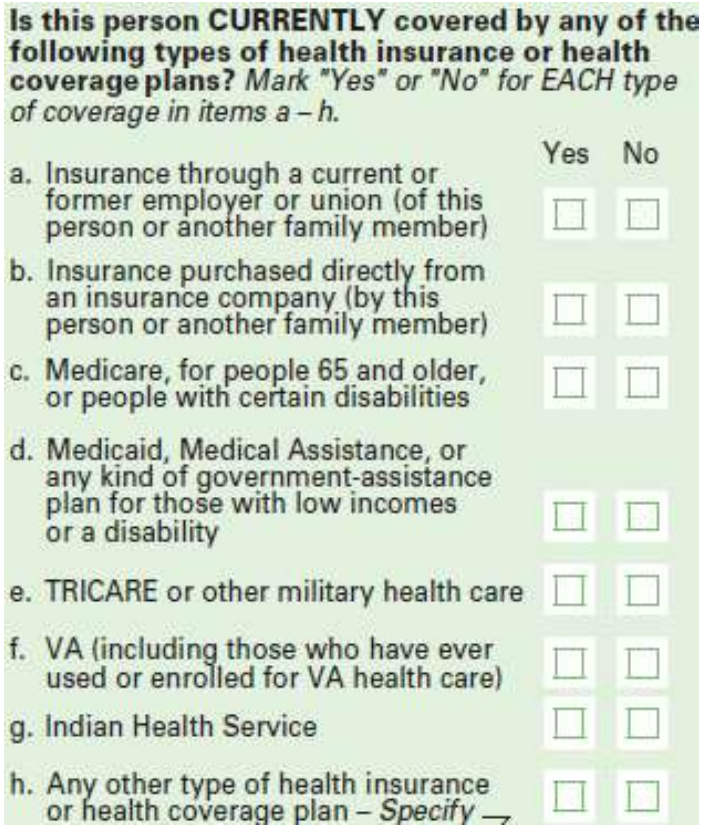
We mostly synthesize variables in descending order of universe size

For variables with the same universe, we synthesize in the order the questions appear on the ACS

Whether variable synthesis order is important is undetermined

Current research focuses on health insurance

Separate variables for types of health insurance appear early in the synthesis



Is this person CURRENTLY covered by any of the following types of health insurance or health coverage plans? *Mark "Yes" or "No" for EACH type of coverage in items a – h.*

	Yes	No
a. Insurance through a current or former employer or union (of this person or another family member)	☐	☐
b. Insurance purchased directly from an insurance company (by this person or another family member)	☐	☐
c. Medicare, for people 65 and older, or people with certain disabilities	☐	☐
d. Medicaid, Medical Assistance, or any kind of government-assistance plan for those with low incomes or a disability	☐	☐
e. TRICARE or other military health care	☐	☐
f. VA (including those who have ever used or enrolled for VA health care)	☐	☐
g. Indian Health Service	☐	☐
h. Any other type of health insurance or health coverage plan – *Specify* →	☐	☐

Health insurance has features we can generalize to other variables

Health insurance is useful to study because

- It consists of multiple underlying variables whose joint distribution is of wide interest
- Health insurance statuses for different members of the household are correlated
- The variables are defined for every ACS respondent

Synthesis methods may take varying approaches to household structure

Method 1: Synthesize each person independently of other members of the household

Method 2:

- Synthesize householders first
- Synthesize other each variable for others using previous variables for that person, corresponding variable for the householder

We determine data quality with a bootstrap simulation

We create a bootstrapped version of each table based on the original data

We compare the L1 distance between the original and synthetic tables with the distance between the original and bootstrapped tables

Values outside the 95% bootstrap confidence interval are asterisked and in red

Private insurance rates change minimally for all people and for householders alone

Private Insurance Rates
Unweighted ACS vs. Synthetic Data

Data	All People	Householders
ACS Unweighted	69.2%	72.2%
Method 1 Synthetic Data	69.3%	72.2%
Method 2 Synthetic Data	68.8%	72.6%

Private insurance rates change noticeably when we subset on relationship to householder

Private Insurance Rates

Unweighted ACS vs. Synthetic Data

Data	Relationship to Householder	
	Spouse	Child
ACS Unweighted	79.8%	64.7%
Method 1 Synthetic Data	80.6%*	64.3%
Method 2 Synthetic Data	74.6%*	65.9%*

* Significantly different from the original data

Neither method preserves complete picture of household insurance rates

Whole-Household Insurance Rate

Unweighted ACS vs. Synthetic Data

Data	% of Households with all members insured
ACS Unweighted	87.0%
Method 1 Synthetic Data	83.7%
Method 2 Synthetic Data	84.7%

* Significantly different from the original data

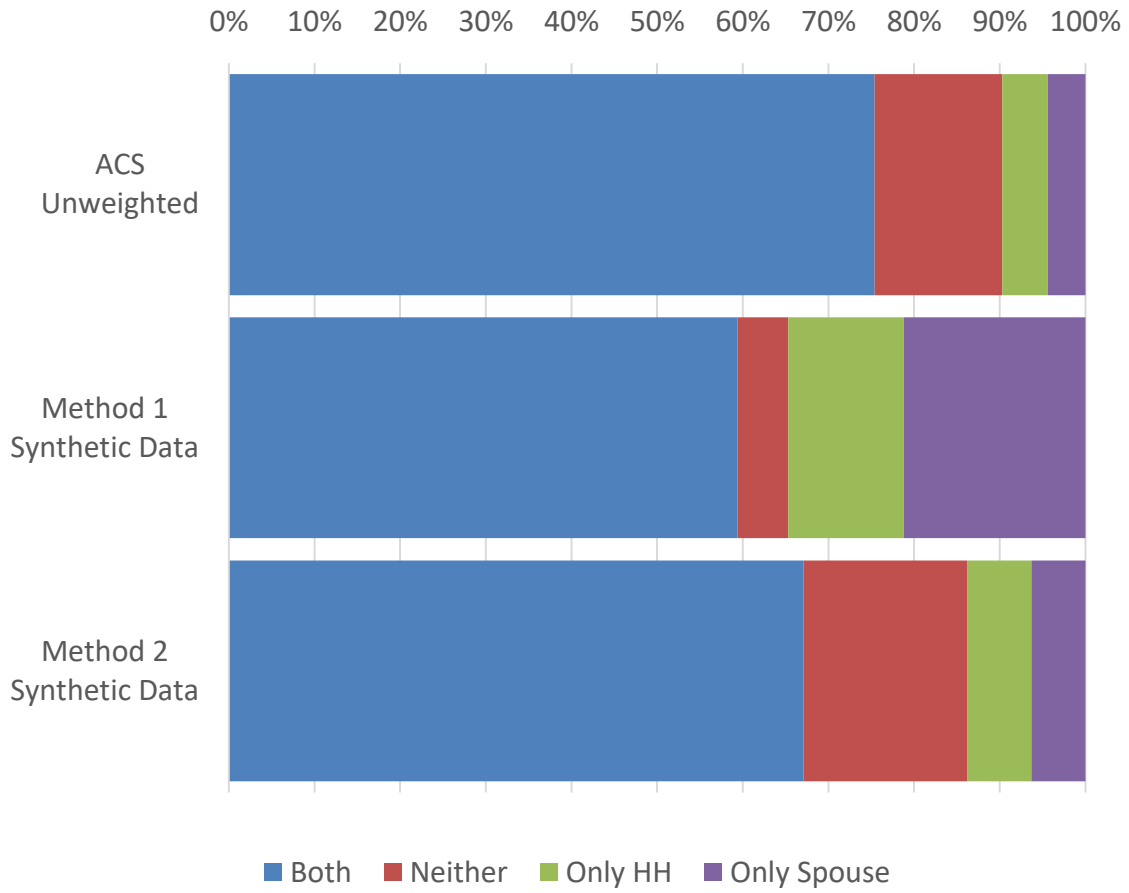
Neither method preserves complete picture of household insurance rates

Both methods change relationship between householder and spouse private insurance statuses

Method 2 is better here, but both methods fall out of the bootstrap range*

Similar results occur if we compare householders' and children's statuses

* Significantly different from the original data



An alternative approach is to generate multiple health insurance statuses simultaneously

Method 1A:

- Create a composite variable that is the cross-tab of all three types of private health insurance (8 categories)
- Generate the composite variable with one tree

Compare one-way, two-way, and three-way marginals of the three-way table depending on whether variables are generated sequentially or simultaneously

Generating health insurance statuses simultaneously has mixed results

One-way, two-way and three-way cross tabs of types of private health insurance

Cross-tab	Within bootstrap variation
Employer	No*
Direct purchase	Yes
TRICARE	Yes
Employer by Direct	No*
Employer by TRICARE	No*
Direct by TRICARE	Yes
Employer by Direct by TRICARE	No*

We seek to improve how the model captures correlations between variables

We have made progress in synthesizing health insurance variables, but we still want to improve our capturing of correlations between variables

The tree method is designed to preserve the strongest relationships in the data, including relationships among more than two variables

Some relationships of smaller magnitude may be important to preserve because of ways the data are used

The path forward presents unresolved challenges

Test data synthesized so far reflect only some of the original data's properties

We need to incorporate weights into the final synthetic data

We need more metrics and benchmarks to assess suitability of various models

We need to research the feasibility of formal privacy for this dataset

Michael Freiman
michael.freiman@census.gov